First steps in signal-processing level models of genetic networks: identifying response pathways and clusters of coexpressed genes

Olga Troyanskaya
lecture for cheme537/cs554

some slides borrowed from Gavin Sherlock, SMD & Stanford Dept of Genetics
Today: focus on microarrays

• A large scale (high-throughput) way to measure gene expression levels on a genomic scale
Why microarray analysis: the questions

- Large-scale study of biological processes
- What is going on in the cell at a certain point in time?
  - what pathways are active, which genes are involved in these pathways
- On the large-scale genomic level, what accounts for differences between phenotypes?
  - which pathways are activated in stress response and which genes belong to these pathways
- Sequence important, but genes have effect through expression
Why is expression important?
Microarray technology
Microarray technologies

• **Spotted cDNA arrays**
  – Developed by Pat Brown (Stanford U)
  – Robotic microspotting
  – PCR products of full-length genes (>100nts)
  – Now also have oligo spotted chips (70bp oligos)

• **Affymetrix GeneChips**
  – Photolithography (from computer industry)
  – Each gene represented by many n-mers – both matches and mismatches

• **Bubble jet / Ink jet arrays**
  – Oligos (25-60 nts) built directly on arrays (in situ synthesis)
  – Highly uniform spots, but expensive
Early cDNA microarray
(18,000 clones)
cDNA microarrays

Known DNA sequences

Isolate mRNA

Glass slide

Cells of Interest

Reference sample

experiments

Resulting data

genes
### Experiments

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cy3</td>
<td>Cy5</td>
<td>log₂(\frac{Cy5}{Cy3})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>10000</td>
<td>50.00</td>
<td>5.64</td>
<td></td>
</tr>
<tr>
<td>4800</td>
<td>4800</td>
<td>1.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>9000</td>
<td>300</td>
<td>0.03</td>
<td>-4.91</td>
<td></td>
</tr>
</tbody>
</table>

Extracting Data
Microarray Data Flow

Microarray experiment → Image Analysis → Database → Data Selection & Missing value estimation → Normalization & Centering → Data Matrix → Supervised Analysis → Decomposition techniques → Networks & Data Integration → Unsupervised Analysis – clustering
Experimental design of microarrays
What can microarrays tell us?

- What genes are involved in specific biological processes (e.g. stress response)
- Assumption = guilt by association (similar expression pattern => same pathway)
- Timing of activation => potential regulation relationships
- Tumor classification for treatment guidance & outcome prediction
Types of experiments

• Time series vs.
• Comparison of groups of samples
• Common reference vs.
• Using reference to compare
Time series

- Measurements taken throughout the time course
- Each array (column of the expression matrix) corresponds to a specific time point
- Can use common reference, or zero-time-point reference
Comparing groups of samples

- Often in clinical studies – can we find similarities or differences within a group of lung cancer patients?
Issues in microarray analysis

Image Analysis

Unsupervised Analysis – clustering

Supervised Analysis

Data Selection & Missing value estimation

Normalization & Centering

Decomposition techniques

Networks & Data Integration

Data Matrix

Database

Microarray experiment
Using reference for comparison

Asynchronous  Synchronized

Prepare RNA
Fluorescently Labeled
cDNA
Hybridize
Common reference
Common reference problem

- Comparison of array experiments from different technologies (even labs) is difficult
- For spotted arrays, data is ratios of sample fluorescence (red) to reference fluorescence (green)
- To compare between experiments, need consistent reference
- “common reference” – a pool of reference mRNA from over 22 cell lines
Microarray analysis at a glance

- **Data Storage & Retrieval**
- Filtering
- Normalization
- Missing value estimation
- Analysis – unsupervised or supervised
- Visualization
Purpose of a microarray DB

Data management

Integration with basic analysis tools

Integration with external information
  consolidation
  data integration

Publication of Results
Example:
Stanford Microarray Database (SMD)

- Data management
  - Storage, archiving and data viewing tools.

- Integration with analysis tools and external information.
  - Clustering, partitioning and output of data for other use. Linkage with SGD and GO.

- Publication of results
  - Provide data, images, analysis and connections with biological resources. Linkage with SGD.
SMD provides:

- Storage of both the raw and normalized data from microarray experiments, as well as their corresponding image files.

- Interfaces for data retrieval, analysis, visualization, and organization.

- A means of associating meaningful information, both biological and methodological, with the experiment. This includes annotation of the arrayed samples, the probe(s), the materials and methods, and the experimental context (groupings).
Scale of the problem by the end of 2001

- 500 slides (experiments) per week
- >40,000 spots per slide
- 1 billion spots/year!
- Uncertain number of organisms to be included.
- 750 GB in TIFF images per year, and growing
Growth of SMD

As of November 27, 2001
SMD Built from Components

- Oracle DBMS
- Web interface via Perl CGI and DBI
- TIFFs and primary data archived to tape and Magneto-optical disks
- GIF pseudocolor images stored outside DBMS
- Microarray data stored in 24 core tables
- External datasets currently in 34 tables
Design challenges, an example

• Need to consider at least two levels of identifier:
  • Physical DNA (SUID) - should track with sequence, though sequence is not always known
  • Genetic Entity to which DNA maps (LocusID)
    • can dynamically change => need regular communication with NIH databases for updating
    • requires that SUID can be easily mapped to the LocusID
• Access issues
Microarray analysis at a glance

• Data Storage & Retrieval
• **Filtering**
• Normalization
• Missing value estimation
• Analysis – unsupervised or supervised
• Visualization
Data Filtering

Goals:
- Extract only experiment/gene subsets of interests
- Extract only “accurate” data points

Various filtering criteria:
- Manual
- Fluorescence distribution
- Level of expression in each channel

Filters can be combined using logical operators
Why worry?  
**Spots with low regression correlation**

Challenge – How can we differentiate between data and noise on image level?
Microarray analysis at a glance

- Data Storage & Retrieval
- Filtering
- **Normalization**
- Missing value estimation
- Analysis – unsupervised or supervised
- Visualization
Data Normalization: Definition

• Normalization is an attempt to compensate for systematic bias in data
• Normalization attempts to remove the impact of non-biological influences on biological data:
  – Balance fluorescent intensities of the two dyes
  – Adjust for differences in experimental conditions (b/w replicate gene expression experiments)
• Normalization allows to compare data from one experiment to another (after removing experiment-specific biases)
Normalization: Sources of Systematic Bias

- Different labeling efficiencies or dye effects (two-channel arrays)
- Scanner malfunction
- Differences in concentration of DNA on arrays (plate effects)
- Printing or tip problems
- Uneven hybridization
- Batch bias
- Experimenter issues
Normalization: Effects on Intensity

Same mRNA hybridized in both channels
Microarray analysis at a glance

- Data Storage & Retrieval
- Filtering
- Normalization
- **Missing value estimation**
- Analysis – unsupervised or supervised
  - Unsupervised: clustering & PCA
  - Supervised
    - Biomarker identification
The missing value problem

• Microarrays can have systematic or random missing values
• Some algorithms can’t deal with missing values
• Large literature on missing data estimation exists
• What’s best to do for microarrays?
Accurate estimation important for analysis

Complete data set

Data set with 30% entries missing and filled with zeros (zero values appear black)

Data set with missing values estimated by KNNimpute algorithm
KNNimpute Algorithm

• Idea: use genes with similar expression profiles to estimate missing values
Microarray analysis at a glance

- Data Storage & Retrieval
- Filtering
- Normalization
- Missing value estimation
- **Analysis – unsupervised or supervised**
  - Unsupervised: clustering & PCA
  - Supervised
    - Biomarker identification
Unsupervised analysis - clustering
Why cluster?

• “Guilt by association” => if unknown gene $i$ is similar in expression to known gene $j$, maybe they are involved in the same/related pathway

• Dimensionality reduction: datasets are too big to be able to get information out without reorganizing the data
What is clustering?

• Reordering of gene (or experiment) expression vectors in the dataset so that similar patterns are next to each other
Clustering Random vs Biological Data

Challenge – when is clustering “real”?

K-means clustering

1. Define $k =$ number of clusters
2. Randomly initialize a seed vector for each cluster
3. Go through all genes, and assign each gene to the cluster which it is most similar to
4. Recalculate all seed vectors as means (or medians) of patterns of each cluster
5. Repeat 3&4 until $<$stop condition$>$
K-means clustering: stop conditions

• Until the change in seed vectors is < <constant>
• Until all genes get assigned to the same partition twice in a row
• Until some minimal number of genes (e.g. 90%) get assigned to the same partition twice in a row
K-means: problems

• Have to set $k$ ahead of time
• Each gene only belongs to 1 cluster
• One cluster has no influence on the others (one dimensional clustering)
• Genes assigned to clusters on the basis of all experiments
Defining \( k \) (# of clusters)

- **Gap statistic**
  - Find \( k \) at which within-cluster variation is min
  - Plot difference between real and random data’s within-cluster variation, choose max difference point

- **Leave-one out cross-validation**
  - Quality of clusters higher if less within-cluster variation on the “test” array

- **Resampling based methods**
Can a gene belong to N clusters?

- Fuzzy clustering: each gene’s relationship to a cluster is probabilistic
- Gene can belong to many clusters
- More biologically realistic, but harder to get to work well/fast
- Harder to interpret
Self Organizing Maps (SOM)

- Similar to k-means
- BUT: allow clusters to influence each other
Self-organizing maps algorithm

1. Partition data (e.g. 3x2 grid)
2. Randomly choose “seed” vectors for each partition (length = # experiments)
3. Pick a gene at random (e.g. gene $i$, see which partition it is most similar to (e.g. partition A), and modify A’s seed vector to be more similar to gene $i$
4. Now modify neighboring partitions of A to be more similar to A
5. After map “settles down”, assign each gene to the most similar partition
1. Initialize the seeds for each partition

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
<td>6 4 5 6 7 8</td>
<td>4 5 6 1 4</td>
</tr>
<tr>
<td>D</td>
<td>1 5 2 3 1</td>
<td>2 3 4 2 3</td>
<td>0 9 0 8 8</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Pick a gene at random, and adjust the closest partition

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Iteration 1.
3. Adjust neighboring partitions

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td>5</td>
<td>R</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>8</td>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Iteration 1.
2. Pick a gene at random, and adjust the closest partition

Iteration 2.
Self-organizing maps iterations

- At higher iterations, smaller $R$
- At higher iterations, smaller change to partition seeds

$\Rightarrow$ the map “settles down”
Self Organizing Maps: Result

- SOMs result in genes being assigned to partitions of most similar genes
- Neighboring partitions are more similar to each other than they are to distant partitions
SOM: problems

• Have to set $n$ and $m$ ahead of time
• Each gene only belongs to 1 cluster
• Genes assigned to clusters on the basis of all experiments
Hierarchical clustering

- Imposes hierarchical structure on all of the data
- Easy visualization of similarities and differences between genes (experiments) and clusters of genes (experiments)
How does Hierarchical Clustering work?

1. Compare all expression patterns to each other.

2. Join patterns that are the most similar out of all patterns.

3. Compare joined patterns to all other un-joined patterns.

4. Go to step 2, and repeat until all patterns are joined.
Hierarchical Clustering
Optimizing node order

• Consider:

  PIR1
  PIR3
  ASH1

• Is Ash1’s expression most similar to Pir1, or Pir3?

• Flip when joining to make most similar patterns adjacent:
Hierarchical clustering: problems

• Hard to define distinct clusters
• Genes assigned to clusters on the basis of all experiments
• Optimizing node ordering hard (finding the optimal solution is NP-hard)
• Can be driven by one strong cluster – a problem for gene expression b/c data in row space is often highly correlated
• Hard to partition into distinct clusters
Choice of distance metric is important

• Treat data for a gene as a vector

• Distance metric important:
  
  • **Linear**: Euclidean distance, or Pearson correlation
  
  • **Nonlinear**: Spearman…

\[
d_{x, y} = \sqrt{\frac{\sum_{j=1}^{n} (x_j - y_j)^2}{n}}
\]

\[
d_{x, y} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{\sigma_x} \right) \left( \frac{y_i - \bar{y}}{\sigma_y} \right)
\]
EVALUATION: Clustering (supervised or unsupervised)

• a new brilliant algorithm is not enough – how does it compare?
• No external standard on real data =>
  – Can use synthetic datasets
  – Beware of assumptions (e.g. normality)
• Internal standards – lots of research in this area!

A difference between a useful bioinformatics advance and a non-relevant publication is most often EVALUATION!
Clustering: Visualization

- Lots of Visualization and HCI challenges:
  - Lots of data
  - Dynamic navigation
  - Simultaneous display of different data types
  - Simultaneous display of different zoom levels for data
  - Dynamic links to other databases

Visualization often critical for late-stage biological analysis! Especially important for biological networks.
Advanced analysis methods – a very brief overview
-decompositions
-supervised methods
Principle Components Analysis

• PCA (or SVD)
• Decomposition technique
• Produces a set of “fundamental” expression patterns = principal components
• A linear combination of principle components can represent the behavior of all of the genes in a given data set
PCA (cont.)

• PCA is not a clustering technique
• A tool to characterize most abundant themes or building blocks that occur in many genes or experiments
• Can use PCA to filter out experimental artifacts
• Could use principal components as cluster seeds to cluster genes
PCA – the math

• Seeks better linear basis to describe the data
  – Max variance orthogonal directions
  – Eigenvalues suggest importance
Supervised methods – a very brief introduction
Supervised methods

• Clustering, PCA etc are unsupervised methods
• Supervised method – any method that “learns” a way to perform an operation based on examples of problems with known solutions (i.e. has “supervision”)
• => have to know what we are looking for when using supervised methods
Supervised vs. Unsupervised

- Unsupervised methods can find novel profile groupings
- Supervised methods take known groupings and create rules for reliably assigning genes or conditions into these groups
Finding ribosomal genes – an example of supervised problem

• Problem: want to classify unknown genes as either ribosomal or non-ribosomal
• Have: a group of known ribosomal genes
• Solution:
  – build a mathematical model (a set of rules) of what a microarray profile of a ribosomal gene looks like
  – use the model to test the unknown genes
Some supervised methods

- Neural networks
- Bayesian networks
- Linear discriminant analysis (LDA)
- Logistic regression
Identifying biomarkers and pathways by response to experimental or natural factors
The problem

• Have samples in two groups A and B
• Want to identify biomarker genes between A and B

• Challenges:
  – Data are not normally distributed
  – Microarray data are often noisy
Why not just look?

- Because even if patterns look promising to us by eye, they may have just happened by chance
- Need to compare our results to some background distribution (random or chance occurrences)
- Could use a known distribution (e.g. normal or binomial) or can generate a background distribution from data
Nonparametric t-test

- Want to pick genes with:
  - Maximal difference in mean expression between samples
  - Minimal variance of expression within sample
Nonparametric t-test

group 1: \(n_1\) samples, with average expression \(\overline{X}_1\)
group 2: \(n_2\) samples, with average expression \(\overline{X}_2\)

t statistic: \(t = \frac{\left(\overline{X}_1 - \overline{X}_2\right)}{S_{\overline{X}_1 - \overline{X}_2}}\)

\[S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\]

p value (from column permutations):

\[p = \frac{\text{count}(t_{j_{\text{perm}}} > t_{j_{\text{obs}}})}{\text{count}(\text{permutations})}\]
Wilcoxon rank-sum test

- Tests for equality of means of two samples
- Uses rank data
- Good for non-normal data

<table>
<thead>
<tr>
<th>Original data</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

- Identifies genes with skewed distribution of ranks
Wilcoxon rank sum test

group 1: \( n_1 \) samples  
group 2: \( n_2 \) samples  (\( n_2 > n_1 \))

\[
\begin{align*}
    w_1 : & \sum \text{ranks}_{\text{sample 1}} \\
    u_1 : & w_1 - n_1 \times (n_1 + 1)/2 \\
    \text{mean}_{u_1} : & n_1 \times n_2 / 2 \\
    \text{var}_{u_1} : & n_1 \times n_2 \times (n_1 + n_2 + 1)/12 \\
    z : & (u_1 - \text{mean}_{u_1}) / \sqrt{\text{var}_{u_1}} \\
    z : & \in N(0,1) \text{ when } n_1 > 8
\end{align*}
\]
Identifying pathways potentially involved in lung cancer progression

• One of the most common cancers in the world

• Four major subtypes: small cell (18%), squamous (30%), adenocarcinoma (30%), large cell (10%)

• Subtypes diagnosed by light microscopy

Garber, Troyanskaya et al. PNAS 2001
Squamous vs. normal lung

- Clones that pass the p-value cutoff of 0.1:
  - RST: 92
  - T test: 202
  - ID: 301
  - In common: 91 clones, or 86 distinct genes

- Of 86 genes identified by all three methods:
  - 48 known proliferation-related genes
  - 10 genes known to be overexpressed in tumors
  - 3 known squamous tumor markers
  - 18 genes with other functions
  - 5 ESTs and hypothetical proteins
Hierarchical clustering of lung cancers

Adeno group 1

Adeno group 2

Normal

Adeno group 3

Squamous

Small cell

Large cell
Patient survival for Adenocarcinoma subgroups

Cum. Survival (Group 1)
Cum. Survival (Group 2)
Cum. Survival (Group 3)

p = 0.002 for Gr. 1 vs. Gr. 3
Identification of Adenocarcinoma subgroups

- Lung adenocarcinomas are heterogeneous in patterns of gene expression.
- Adenocarcinoma tumors are subdivided into 3 distinct subgroups by hierarchical clustering.
- Adenocarcinoma subgroups correlate with survival.
- Can we identify individual genes that can predict patient survival for adenocarcinoma lung cancer? Do these genes point to a specific pathway?
Genes differentially expressed in Adenocarcinoma subgroups

- Applied nonparametric t-test to lung cancer dataset (Garber, Troyanskaya et al. 2001)
- Marker genes identified seem promising from current antibody staining results
- Genes overexpressed in low survival Adenocarcinomas

<table>
<thead>
<tr>
<th>90 genes total</th>
</tr>
</thead>
<tbody>
<tr>
<td>solute carrier family 7, member 5 (CD98)</td>
</tr>
<tr>
<td>ataxia-telangiectasia D-associated EST, 251278</td>
</tr>
<tr>
<td>prostaglandin E synthase EST, 11607</td>
</tr>
<tr>
<td>cathepsin L</td>
</tr>
<tr>
<td>dickkopf homolog 1</td>
</tr>
<tr>
<td>LTB4-12 hydroxydehydrogenase EST, Hs.135056</td>
</tr>
<tr>
<td>vascular endothelial growth factor C EST, Hs.135056</td>
</tr>
<tr>
<td>cDNA DKFZp761G02121</td>
</tr>
</tbody>
</table>
Application to a lymphoma dataset

- Data from Alizadeh et al. (2000)
  - Two molecular subtypes of DLBCL
  - Clinically heterogeneous
- RST identified 72 clones (70 unique sequences)
- Perfect separation by clustering based on 72 clones identified
- 57 genes in common with Alizadeh et al. list of over 350 clones
- 13 new differentially expressed clones identified
A

B

BCL-6
BCL-7A
BM1
CD10
CBF/PEBP2aA1/AML1
Cyclin D2
Cyclin H
Deoxycytidylyate deaminase
DNA (cytosine-C)-methyltransferase
FLICE-like inhibitory protein long form (I-FLICE)
FMR2=Frail X mental retardation 2
FR4
JAK1 lymphoid-restricted membrane protein
JNK3 Stress-activated protein kinase
MCL1 myeloid cell differentiation protein
myc-related gene A=myc gene
PAK-beta
Potassium voltage-gated channel shaker-related mem 3
PRK putative serine/threonine protein kinase
PTP-1B phosphotyrosyl-protein phosphatase
RPS3 homologue of yeast RPS3 transcription factor
SLAP src-like adapter protein
T-cell protein-tyrosine phosphatase
TdT terminal deoxynucleotidyl transferase
TIG-2/Rhombotin-2
zinc finger protein 42 MZF-1
Analysis summary

• Unsupervised methods –
  – clustering
    • Hierarchical clustering
    • Self organizing maps
    • K means clustering
  – Decomposition
    • Principle components analysis

• Supervised methods
  – Require examples with known answers
  – Need both positive and negative examples

• Biomarker identification
  – Nonparametric t-test
  – Rank sum test
Back to networks

- Groups of coexpressed genes =? Pathway candidates
- By strategically choosing conditions and looking for differentially expressed genes, can identify pathway players
- Time series experiments imply regulation cascades
- Incorporating other types of data increases accuracy and clarifies the picture
End of lecture

Thank you

ogt@princeton.edu